SSOAR

Open Access Repository
www.ssoar.info

Regressionsanalyse mit Panel-Daten: eine Einführung

Alecke, Björn

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Alecke, B. (1997). Regressionsanalyse mit Panel-Daten: eine Einführung. *ZA-Information / Zentralarchiv für Empirische Sozialforschung*, 40, 87-121. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-200295

Nutzungsbedingungen:
Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.
By using this particular document, you accept the above-stated conditions of use.

gesis Leibniz-Institut für Sozialwissenschaften

Mitglied der Leibniz-Gemeinschaft

Regressionsanalyse mit Panel-Daten: Eine Einführung

von Björn Alecke[1]

Zusammenfassung

Dieser Beitrag gibt eine Einführung in die verschiedenen Verfahren zur regressionsanalytischen Behandlung von Panel-Daten, in der auf Ableitungen und eine extensive Benutzung von Matrix-Algebra verzichtet wird. Allerdings kann die Darstellung nicht ganz ohne eine formale Schreibweise auskommen, wobei jedoch im letzten Abschnitt anhand eines konkreten Rechenbeispiels die benutzten Formeln näher erläutert werden. Auf diese Weise möchte der Beitrag einerseits aufzeigen, daß die Regressionsanalyse mit Panel-Daten im wesentlichen nur auf rechentechnisch einfach durchzuführende Transformationen der Daten hinausläuft und mit Hilfe der üblichen Statistik-Programmpakete (z.B. SPSS) durchgeführt werden kann, und andererseits einen leichteren Zugang zur bestehenden Lehrbuchliteratur ermöglichen.

Abstract

This article gives a short introduction into existing methods of investigating panel data by means of regression analysis without relying on extensive use of matrix algebra and formal derivatives. Although a formal presentation can not be completely avoided, a simple example is given in the final section to illustrate the main formulas. By doing this, the article is intended on the one hand, to show that regression analysis of panel data requires in essence only straightforward transformations of data, and on the other hand, to permit a more readily accessability to the textbook literature for interested readers.

[1] ***Björn Alecke*** (Dipl.-Vw.) ist wissenschaftlicher Mitarbeiter an der Universität Münster, Institut für Wirtschafts- und Sozialgeschichte, Hüfferstr. 1a, 48149 Münster.

I. Einleitung

Mit Panel- oder auch Longitudinaldaten bezeichnet man einen Datensatz, der bei $i = 1,2,\ldots N$ Untersuchungseinheiten (Merkmalsträgern) die beobachteten Werte einer oder mehrerer Variablen (die Merkmale) für $t = 1,2,\ldots T$ verschiedene Zeitpunkte erfaßt. Als Beispiel für einen Panel-Datensatz könnte man etwa die Erfassung von Stimmenanteilen für politische Parteien und bestimmte sozio-ökonomischen Variablen bei *N* Wahlkreisen über *T* Wahlperioden anführen. Möchte man mit Hilfe dieses Datensatzes beispielsweise die Hypothese überprüfen, daß der Anteil der SPD-Wähler vom Ausmaß der Arbeitslosigkeit beeinflußt wird, so könnte man folgende Regressionsfunktion $SPD = \alpha + \beta \cdot ALQ$ schätzen, wobei die Verwendung von Paneldaten gegenüber reinen Querschnitts- bzw. Zeitreihendaten eine Reihe von Vorteilen bietet: unmittelbar offensichtlich ist die gestiegene Zahl von Freiheitsgraden, da die Stichprobengröße hier *NT* beträgt. Dies wird die Genauigkeit (Effizienz) der Schätzung erhöhen. Ebenso vermindert sich die Gefahr von Multikollinearität, da im allgemeinen bei Paneldaten die Streuung zwischen den erklärenden Variablen größer sein wird. Ein wesentlicher Vorteil liegt darin, daß erst Paneldaten die Beantwortung bestimmter ökonomischer Fragestellungen ermöglichen. Dazu sei ein von ***Baltagi*** (1995, S.5) angeführtes Beispiel wiedergegeben:

"Suppose that we have a cross-section of women with a 50% average yearly labour force participation rate. This might be due to (a) each woman having a 50% chance of being in the labour force, in any given year, or (b) 50 % of the women work all the time and 50% do not. Case (a) has high turnover, while case (b) has no turnover. Only panel data could discriminate between these cases".

Der zusätzliche Nutzen von Paneldaten ist jedoch auch mit Kosten verbunden, die in der Form höherer Anforderungen bei der Durchführung der Regressionsanalyse bestehen. Deutlich wird dies in der Vielzahl der in der Ökonometrie entwickelten Verfahren zur Behandlung von Paneldaten, die in der untenstehenden Tabelle 1 aufgeführt werden.

Die Verfahren unterscheiden sich dabei hinsichtlich der getroffenen Annahmen über den deterministischen und stochastischen Teil des (linearen) Regressionsmodells. Unterschiedliche Annahmen über den stochastischen Teil sind auch aus der herkömmlichen Regressionsanalyse bekannt. Zumeist wird davon ausgegangen, daß bei Zeitreihenuntersuchungen eine Autokorrelation der Residuen bestehen kann, während bei Querschnittsuntersuchungen oftmals Heteroskedastizität der Residuen zu beobachten ist. Bei Paneldaten als kombinierte Querschnitts - und Zeitreihenuntersuchung können deshalb in vielen Fällen Residuen vermutet werden, die sowohl durch Autokorrelation als auch durch Heteroskedastizität gekennzeichnet sind.

Tabelle 1: Taxonomie von Regressionsmodellen mit kombinierten Zeitreihen- und Querschnittsdaten

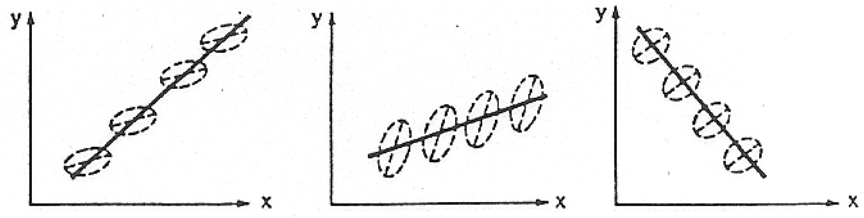
Annahmen über			
Regressionskonstante	Regressionsgewichte	Residuen	Modellbezeichnung
gemeinsam über alle *i* und *t*	gemeinsam über alle *i* und *t*	$E(\mathbf{ee}') = \sigma_e^2 \mathbf{I}_{NT}$	Classical Pooling
gemeinsam über alle *i* und *t*	gemeinsam über alle *i* und *t*	$E(\mathbf{ee}') = \mathbf{V}$	Kmenta-Model
verschieden über alle *i*	gemeinsam über alle *i* und *t*	Fixed effects	Least Squares Dummy Variable-Model
verschieden über alle *i*	gemeinsam über alle *i* und *t*	Random effects	Error Components Model
verschieden über alle *i* und *t*	gemeinsam über alle *i* und *t*	Fixed effects	Least Squares Dummy Variable-Model
verschieden über alle *i* und *t*	gemeinsam über alle *i* und *t*	Random effects	Error Components Model
verschieden über alle *i*	verschieden über alle *i*	Fixed effects	SURE Model
verschieden über alle *i*	verschieden über alle *i*	Random effects	Swamy Random Coefficient Model
verschieden über alle *i* und *t*	verschieden über alle *i* und *t*	Random effects	Hsiao Random Coefficient Model

Quelle: vgl. Johnston (1986), S. 397 und Judge (1985), S. 517.

Die verschiedenen Annahmen über den deterministischen Teil, also divergierende Parameterwerte für verschiedene Individuen und Zeitpunkte, lassen sich mit einer eventuell bestehenden Heterogenität von Paneldaten begründen, deren Nichterkennen zu Verzerrungen bei der Koeffizientenschätzung führen kann (Heterogenitätsbias). In Anlehnung an ***Hsiao*** (1986, S.6) sei folgendes Beispiel für die Schätzung der obigen Regressionsfunk-tion zwischen Stimmenanteil und Arbeitslosigkeit gegeben. Angenommen, man habe für den gesamten Datensatz, d.h. über alle Wahlkreise und -perioden, die Regressionsfunktion ge-

schätzt und dabei unterstellt, die Parameterwerte für α und β seien für alle Wahlkreise gleich, so könnten sich die in den Abbildungen 1 und 2 dargestellten Fälle ergeben.

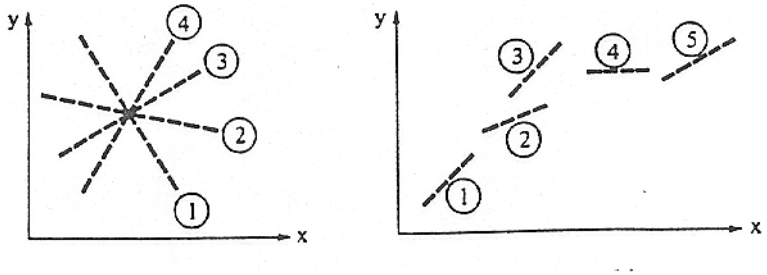
Abbildung 1



1a) 1b) 1c)

Quelle:Hsiao (1986), S.7

In Abbildung 1 wurde angenommen, daß jeder der (hier zur Vereinfachung nur 4 dargestellten) Wahlkreise einen anderen Wert für die Regressionskonstante α aufweise, so daß eigentlich die gestrichelten Linien die jeweiligen für die Wahlkreise gültigen Regressionsfunktionen angeben sollten. Vernachlässigt man den individuellen Unterschied in den Regressionskonstanten, so wird die durchgezogene Linie geschätzt, die offensichtlich sowohl für die Regressionskonstante als auch für das Regressionsgewicht β einen falschen Schätzwert liefert. Dabei zeigt sich, daß je nach Situation die Richtung der Verzerrungen unterschiedlich sein kann (Fall 1a) Überschätzung von β, 1b) Unterschätzung, 1c) falsches Vorzeichen). **Abbildung 2**



2a) 2b)

Quelle: Hsiao (1986), S.7

Abbildung 2 zeigt den Fall, daß sowohl α und β für die Wahlkreise verschieden sind. Im Fall 2a) würde die (nicht gezeigte) Regressionsgerade für den kompletten Datensatz eine Art Mittelwert aus den individuellen Regressionsgeraden bilden, während für den Fall 2b) sich eine logarithmische Form ergäbe. Folglich muß man bei Paneldaten in Erwägung ziehen, daß die zu schätzenden Parameter über die Individuen und/oder über die Zeit variieren.

Grundsätzlich lassen sich hier vier Fälle unterscheiden:

1. Die Regressionsgewichte sind konstant, aber die Regressionskonstante variiert über die Individuen:

$$y_{it} = \beta_{1,i} + \beta_2 x_{2,it} + \beta_3 x_{3,it} + \ldots + \beta_K x_{K,it} + e_{it}$$

2. Die Regressionsgewichte sind konstant, aber die Regressionskonstante variiert über die Individuen und die Zeit:

$$y_{it} = \beta_{1,it} + \beta_2 x_{2,it} + \beta_3 x_{3,it} + \ldots + \beta_K x_{K,it} + e_{it}$$

3. Alle Koeffizienten, also Regressionsgewichte und Regressionskonstante, variieren über die Individuen:

$$y_{it} = \beta_{1,i} + \beta_{2,i} x_{2,it} + \beta_{3,i} x_{3,it} + \ldots + \beta_{K,i} x_{K,it} + e_{it}$$

4. Alle Koeffizienten, also Regressionsgewichte und Regressionskonstante, variieren über die Individuen und die Zeit:

$$y_{it} = \beta_{1,it} + \beta_{2,it} x_{2,it} + \beta_{3,it} x_{3,it} + \ldots + \beta_{K,it} x_{K,it} + e_{it}$$

Diese Fallunterscheidungen spiegeln sich in Tabelle 1 in den unterschiedlichen Annahmen über den deterministischen Teil der zu schätzenden Regressionsfunktion wider. In den folgenden Abschnitten werden die angeführten Modelle näher vorgestellt, wobei im letzten Abschnitt ein simplifiziertes Rechenbeispiel gegeben wird.

II. Modelle mit konstantem Parametervektor

1. Das "Classical Pooling"-Modell

Dieses Modell ist eine einfache Erweiterung des klassischen (linearen) Regressionsverfahrens auf einen Paneldatensatz. Neben dem konstanten Parametervektor für jedes Individuum und über alle Zeitpunkte wird für die Residuen Homoskedastiziät und fehlende Autokorrelation sowie auch eine fehlende Korrelation ihrer Ausprägungen zwischen den Individuen unterstellt.

Für den Fall von *K* erklärenden Variablen lauten die Beobachtungsgleichungen für *N* Individuen ($i=1,2,..N$) und bei *T* Zeitpunkten ($t=1,2...T$):

$$
\begin{aligned}
y_{11} &= \beta_1 + \beta_2 x_{2,11} + \beta_3 x_{3,11} + \ldots + \beta_K x_{K,11} + e_{11} \\
y_{12} &= \beta_1 + \beta_2 x_{2,12} + \beta_3 x_{3,12} + \ldots + \beta_K x_{K,11} + e_{12} \\
&\vdots \\
y_{1T} &= \beta_1 + \beta_2 x_{2,1T} + \beta_3 x_{3,1T} + \ldots + \beta_K x_{K,1T} + e_{1T} \\
y_{21} &= \beta_1 + \beta_2 x_{2,21} + \beta_3 x_{3,21} + \ldots + \beta_K x_{K,21} + e_{21} \\
y_{22} &= \beta_1 + \beta_2 x_{2,22} + \beta_3 x_{3,22} + \ldots + \beta_K x_{K,22} + e_{22} \\
&\vdots \\
y_{2T} &= \beta_1 + \beta_2 x_{2,2T} + \beta_3 x_{3,2T} + \ldots + \beta_K x_{K,2T} + e_{2T} \\
&\vdots \\
y_{N1} &= \beta_1 + \beta_2 x_{2,N1} + \beta_3 x_{3,N1} + \ldots + \beta_K x_{K,N1} + e_{N1} \\
y_{N2} &= \beta_1 + \beta_2 x_{2,N2} + \beta_3 x_{3,N2} + \ldots + \beta_K x_{K,N2} + e_{N2} \\
&\vdots \\
y_{NT} &= \beta_1 + \beta_2 x_{2,NT} + \beta_3 x_{3,NT} + \ldots + \beta_K x_{K,NT} + e_{NT}
\end{aligned}
$$

Das Gleichungssystem, bestehend aus insgesamt *NT* einzelnen Beobachungsgleichungen, läßt sich kompakt in Matrixnotation schreiben

$$\mathbf{y} = \mathbf{X}ß + \mathbf{e}$$

wobei

$$
\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1T} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{NT} \end{bmatrix}_{NT \times 1}
\qquad
\mathbf{X} = \begin{bmatrix} 1 & x_{2,11} & x_{3,11} & \cdots & x_{K,11} \\ 1 & x_{2,12} & x_{3,12} & \cdots & x_{K,12} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2,1T} & x_{3,1T} & \cdots & x_{K,1T} \\ 1 & x_{2,21} & x_{3,21} & \cdots & x_{K,21} \\ 1 & x_{2,22} & x_{3,22} & \cdots & x_{K,22} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2,NT} & x_{3,NT} & \cdots & x_{K,NT} \end{bmatrix}_{NT \times K}
\qquad
\mathbf{e} = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1T} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{NT} \end{bmatrix}_{NT \times 1}
$$

und

$$ß = (\beta_1 \beta_2 \cdots \beta_K)'$$

Die Varianz-Kovarianzmatrix der Residuen **V** läßt sich allgemein darstellen als

$$\mathbf{V} = \begin{bmatrix} E(e_{11}^2) & E(e_{11}e_{12}) & \cdots & E(e_{11}e_{1T}) & E(e_{11}e_{21}) & E(e_{11}e_{22}) & \cdots & E(e_{11}e_{NT}) \\ E(e_{12}e_{11}) & E(e_{12}^2) & \cdots & E(e_{12}e_{1T}) & E(e_{12}e_{21}) & E(e_{12}e_{22}) & \cdots & E(e_{12}e_{NT}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ E(e_{1T}e_{11}) & E(e_{1T}e_{12}) & \cdots & E(e_{1T}^2) & E(e_{1T}e_{21}) & E(e_{1T}e_{22}) & \cdots & E(e_{1T}e_{NT}) \\ E(e_{21}e_{11}) & E(e_{21}e_{12}) & \cdots & E(e_{21}e_{1T}) & E(e_{21}^2) & E(e_{21}e_{22}) & \cdots & E(e_{21}e_{NT}) \\ E(e_{22}e_{11}) & E(e_{22}e_{12}) & \cdots & E(e_{22}e_{1T}) & E(e_{22}e_{21}) & E(e_{22}^2) & \cdots & E(e_{22}e_{NT}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ E(e_{NT}e_{11}) & E(e_{NT}e_{12}) & \cdots & E(e_{NT}e_{1T}) & E(e_{NT}e_{21}) & E(e_{NT}e_{22}) & \cdots & E(e_{NT}^2) \end{bmatrix}_{NT\times NT}$$

wobei die Blöcke entlang der Hauptdiagonalen die Varianz-Kovarianzmatrix für das jeweilige Individuum angeben, während abseits der Diagonalen die Kovarianzen abgebildet werden, die zwischen den Individuen bestehen. Die Matrix hat also folgende Struktur

$$\begin{bmatrix} [\mathbf{V_1}] & [\mathbf{V_{1,2}}] & \cdots & [\mathbf{V_{1,N}}] \\ [\mathbf{V_{2,1}}] & [\mathbf{V_2}] & \cdots & [\mathbf{V_{2,N}}] \\ \vdots & \vdots & \vdots & \vdots \\ [\mathbf{V_{N,1}}] & [\mathbf{V_{N,2}}] & \cdots & [\mathbf{V_N}] \end{bmatrix}$$

Unter den Annahmen des klassischen Regressionsmodells von Homoskedastizität

$E(e_{it}^2) = \sigma_e^2$ für $i = 1,2,\ldots N$

und fehlender Korrelation der Residuen zu verschiedenen Zeitpunkten und zwischen den Individuen

$E(e_{it}e_{is}) = 0 \quad (t \neq s)$

$E(e_{it}e_{js}) = 0 \quad (i \neq j)$ für $t,s = 1,2,\ldots T$

läßt sie sich einfacher als

$$\mathbf{V} = \begin{bmatrix} \sigma_e^2 & 0 & \cdots & 0 \\ 0 & \sigma_e^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_e^2 \end{bmatrix}_{NT\times NT} = \sigma_e^2 \mathbf{I}_{NT}$$

mit $\mathbf{I}_{NT}$ als Einheitsmatrix der Ordnung *NT* schreiben. Weil zwischen den Individuen keine Korrelation besteht, weisen die Blöcke $\mathbf{V_{i,j}}$ abseits der Hauptdiagonalen den Wert 0 auf. Wegen der Unabhängigkeitsannahme der Residuen zu verschiedenen Zeitpunkten für jedes

einzelne Individuum sind innerhalb der Blöcke $\mathbf{V_1}, \mathbf{V_2}$ ... $\mathbf{V_N}$ nur die Elemente auf der Diagonalen (d.h. zu gleichen Zeitpunkten) besetzt. Da für jedes Individuum diese Varianz σ_e^2 als gleich angenommen wurde, lassen sich diese Blöcke als N Einheitsmatrizen der Ordnung T interpretieren, so daß sich insgesamt die Matrix wie oben aufgezeigt ergibt.

Vorausgesetzt die getroffenen Annahmen stimmen mit dem datengenerierenden Prozeß überein, kann dieses Modell mit der üblichen Kleinste Quadrate Methode (Ordinary Least Squares, im folgenden OLS) geschätzt werden. Unterstellt man weiterhin, wie üblich, normalverteilte Residuen, können die bekannten Verfahren zur Hypothesenüberprüfung und zur Bildung von Konfidenzintervallen verwendet werden.

2. Das "Kmenta"-Modell

Wie weiter oben schon erwähnt kann nicht immer davon ausgegangen werden, daß die restriktiven Annahmen über die Residuen erfüllt sind. ***Kmenta*** (1986, S.618ff.) hat deshalb die Schätzung des "Classical Pooling"-Modells dahingehend modifiziert, daß Heteroskedastizität und Autokorrelation berücksichtigt werden. Es ändert sich folglich die Struktur der Varianz-Kovarianzmatrix der Residuen.

Es besteht jetzt die Möglichkeit zu einer individuell verschiedenen Varianz der Residuen:

$$E(e_{it}^2) = \sigma_i^2$$

Die Residuen für jedes Individuum sind autokorreliert:

$$e_{it} = \rho_i e_{i,t-1} + u_{it}$$

wobei $u_{it} \sim N(0, \sigma_{ui}^2)$, $e_{it} \sim N(0, \frac{\sigma_{ui}^2}{1-\rho_i^2})$ und $E(e_{i,t-1}u_{jt}) = 0$ für alle i,j

Zu beachten ist, daß hier die Möglichkeit verschiedener Autokorrelationskoeffizienten für die einzelnen Individuen gegeben ist. Aus obigen Annahmen läßt sich die Kovarianz der Residuen des jeweiligen Individuums zu verschiedenen Zeitpunkten ableiten:

$$\mathrm{E}\left(\mathrm{e_{it}}\,\mathrm{e_{is}}\right) = \rho_{\mathrm{i}}^{\mathrm{t-s}}\sigma_{\mathrm{i}}^2 \quad (\mathrm{t} \neq \mathrm{s})$$

Weiterhin wird jedoch angenommen, daß zwischen den Individuen keine Korrelation besteht,

$$E(e_{it}e_{js}) = 0 \quad (i \neq j) \text{ für } t,s = 1,2,\ldots T$$

so daß sich durch Einsetzen dieser Werte in die allgemeine Form der Varianz-Kovarianzmatrix der Residuen **V**, eine blockdiagonale Darstellung für diese ergibt

$$\mathbf{V} = \begin{bmatrix} \mathbf{V_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{V_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{V_N} \end{bmatrix}_{NT \times NT}$$

wobei

$$\mathbf{V_i} = \sigma_i^2 \begin{bmatrix} 1 & \rho_i & \rho_i^2 & \cdots & \rho_i^{T-1} \\ \rho_i & 1 & \rho_i & \cdots & \rho_i^{T-2} \\ \rho_i^2 & \rho_i & 1 & \cdots & \rho_i^{T-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_i^{T-1} & \rho_i^{T-2} & \rho_i^{T-3} & \cdots & 1 \end{bmatrix}_{T \times T}$$

und jede **0** für eine Matrix der Ordnung [*T*x*T*] steht.

Da die Annahmen des klassischen Regressionsmodells verletzt sind, d.h. die Varianz-Kovarianzmatrix der Residuen besitzt nicht mehr die Form $\mathbf{V} = \sigma_e^2 \mathbf{I}_{NT}$, würde eine Schätzung dieses Modells mit Hilfe der OLS Methode zwar zu konsistenten, aber nicht mehr effizienten Schätzern führen. Hinzu kommt, daß die OLS-Formel für die Berechnung der Standardfehler des Parametervektors verzerrt ist, so daß im folgenden das Verallgemeinerte KQ-Verfahren (Generalized Least Squares, GLS) anzuwenden ist. Allerdings wird theoretisch davon ausgegangen, die Werte der Matrix **V** seien bekannt, man hätte also die nötigen Informationen über die Varianzen und Kovarianzen der nicht beobachtbaren Residuen. Der Parametervektor läßt sich beim GLS-Verfahren über die Formel

$$\tilde{\boldsymbol{ß}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

schätzen. Die Verwendung der Formel ist aber nicht operational, da die meisten Computer-Programme eine Schätzung des Parametervektors über die OLS-Formel vornehmen und die GLS-Formel nicht implementiert ist. Nun läßt sich aber zeigen, daß unter allgemeinen Voraussetzungen die Matrix **V** sich folgendermaßen zerlegen läßt.

$$\mathbf{PVP}' = \sigma_e^2 \mathbf{I}.$$

Multipliziert man die Beobachtungsgleichung auf beiden Seiten mit der Matrix **P,** so ergibt sich

$$\mathbf{Py} = \mathbf{PX}\boldsymbol{ß} + \mathbf{Pe} \text{ bzw.} \mathbf{y}^* = \mathbf{X}^*\boldsymbol{ß} + \mathbf{e}^*$$

Für die Varianz-Kovarianzmatrix der Residuen **e*** des transformierten Modells sind nun die klassischen Annahmen wieder erfüllt, wie aus

$$E[\mathbf{e}^{*}\mathbf{e}^{*\prime}] = E[\mathbf{P}\mathbf{e}\mathbf{e}'\mathbf{P}'] = \mathbf{P}E[\mathbf{e}\mathbf{e}']\mathbf{P}' = \mathbf{P}\mathbf{V}\mathbf{P}' = \sigma_e^{\,2}\mathbf{I}$$

ersichtlich ist. Wendet man das OLS-Verfahren auf das transformierte Modell an, so ergibt sich der gesuchte GLS-Schätzer $\tilde{\boldsymbol{ß}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$.

In der Praxis sind allerdings die Varianzen und Kovarianzen der nicht beobachtbaren Residuen, also die Werte von **V**, nicht bekannt. Die Lösung dieses Problems liegt in einer Schätzung dieser Werte und der entsprechenden Transformationsmatrix, die dann zur Berechnung von **ß** verwandt werden. Dieses Verfahren bezeichnet man deshalb mit Estimated Generalized Least Squares (EGLS) mit der entsprechenden Formel

$$\hat{\tilde{\boldsymbol{ß}}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$$

Dabei läßt sich zeigen, daß eine konsistente Schätzung der Werte in $\hat{\mathbf{V}}$ zu einem Schätzwert von **ß** führt, der in großen Stichproben dieselben optimalen Eigenschaften des GLS-Schätzers aufweist. Ein bekanntes Beispiel für eine EGLS-Schätzung ist das sogenannte Cochrane-Orcutt Verfahren zur Beseitigung von Autokorrelation. Genau dieses Verfahren wird auch von ***Kmenta*** zur Ermittlung einer Transformationsmatrix übernommen, mit der eine Beseitigung der Autokorrelation der Residuen im ersten Schritt erfolgt. Hierauf folgt dann in einem zweiten Schritt eine weitere Transformation der beobachteten Werte zur Beseitigung der Heteroskedastizität. Im einzelnen lauten die Schritte:

- Durchführung einer OLS-Schätzung für alle NT Beobachtungswerte und anschließende Berechnung des Autokorrelationskoeffizienten für jedes Individuum über folgende Formel:

$$\hat{\rho}_i = \frac{\sum_{t=2}^{T} e_{it}e_{i,t-1}}{\sum_{t=2}^{T} e_{i,t-1}^2}$$

- Transformation der Beobachtungswerte von erklärter und erklärenden Variablen gemäß

$$\begin{aligned}
y_{it}^{*} &= \sqrt{1-\hat{\rho}_i^2}\; y_{it} && \textbf{für } t = 1\\
y_{it}^{*} &= y_{it} - \hat{\rho}_i y_{i,t-1} && \textbf{für } t = 2,3\ldots T\\
x_{k,it}^{*} &= \sqrt{1-\hat{\rho}_i^2}\; x_{k,it} && \textbf{für } t = 1\\
x_{k,it}^{*} &= x_{k,it} - \hat{\rho}_i x_{k,i,t-1} && \textbf{für } t = 2,3\ldots T
\end{aligned}$$

OLS-Schätzung über das folgende transformierte Modell

$$y_{it}^{*} = \beta_1^{*} + \beta_2 x_{2,it}^{*} + \beta_3 x_{3,it}^{*} + \ldots + \beta_K x_{K,it}^{*} + e_{it}^{*}$$

- Anschließende Schätzung der Varianz der Residuen e_{it} * für jedes Individuum über

$$\hat{\sigma}_i^2 * = \frac{1}{T-K} \sum_{t=1}^{T} \hat{e}_{it}^{*2}$$

wobei diese Varianzschätzung zur weiteren Transformation verwendet wird, um die noch im Ansatz befindliche Heteroskedastiztät zu eliminieren.

- Nochmalige Transformation von y_{it} * und $x_{k,it}$ * gemäß

$$y_{it} ** = \frac{y_{it}*}{\hat{\sigma}_i *} \text{ und } x_{k,it} ** = \frac{x_{k,it}*}{\hat{\sigma}_i *}$$

und OLS-Schätzung folgender Modellgleichung

$$y_{it} ** = \beta_1 ** + \beta_2 x_{2,it} ** + \beta_3 x_{3,it} ** + \ldots + \beta_K x_{K,it} ** + e_{it} **$$

Da die Residuen e_{it} ** nach diesen Transformationen wieder die klassischen Annahmen erfüllen, also weder Autokorrelation noch Heteroskedastizität aufweisen, besitzt die OLS-Schätzung von **ß** in großen Stichproben wieder wünschenswerte Eigenschaften und die üblichen Verfahren zur Hypothesenüberprüfung und Bildung von Konfidenzintervallen können angewandt werden.

Eine Vereinfachung des obigen Modells besteht in der Annahme eines für alle Individuen gleichen Autokorrelationskoeffizienten, der über die Formel

$$\hat{\rho} = \frac{\sum_{i=1}^{N} \sum_{t=2}^{T} e_{it} e_{i,t-1}}{\sum_{i=1}^{N} \sum_{t=2}^{T} e_{i,t-1}^2}$$

konsistent geschätzt werden kann. Diese Annahme kann insbesondere bei Datensätzen mit wenigen Zeitreihenbeobachtungen angebracht sein, da selbst bei voneinander abweichenden Autokorrelationskoeffizienten der Individuen durch diese Vereinfachung ein Effizienzgewinn erzielt werden kann, der sich aufgrund des größeren Unsicherheitsbereichs einer Schätzung individuell verschiedener Autokorrelationskoeffizienten auf der Basis nur weniger Zeitreihenbeobachtungen ergibt.

Bevor jedoch die obigen Transformationen im Rahmen des "***Kmenta***"-Modells vorgenommen werden, empfiehlt sich auf der Grundlage der berechneten Werte für die Autokorrelation und Varianz der Residuen eine Überprüfung der Hypothese,daß tatsächlich eineEGLS-Schätzung vonnöten ist. Hierzu können die verschiedenen in der ökonometrischen Literatur vorgeschlagenen formalen Testprozeduren auf Autokorrelation (beispielsweise der ***Durbin***-

Watson Test) oder Heteroskedastizität (beispielsweise der ***Goldfeld-Quandt*** Test) herangezogen werden (siehe hierzu etwa ***Judge*** et al. (1989), Kap.9).

Eine wesentliche Erweiterung des "***Kmenta***"-Modells besteht in der Annahme einer möglichen Korrelation der Residuen zwischen den Individuen jeweils zu gleichen Zeitpunkten ("contemporaneous correlation"),

$$E(e_{it}e_{jt}) = \sigma_{ij} \neq 0 \quad (i \neq j)$$

so daß die Matrix **V** nicht mehr blockdiagonal ist. Auch hier kann das EGLS-Verfahren über eine konsistente Schätzung der Elemente von **V** durchgeführt werden (siehe hierzu ***Kmenta*** (1986), S. 622ff.).

III. Modelle mit variablen Regressionskonstanten

1. Individual- und Zeiteffekte

Bisher wurde davon ausgegangen, daß für alle Individuen zu jedem Zeitpunkt der Parametervektor konstant sei. Weiter oben wurde darauf hingewiesen, daß eine solche Annahme in vielen Fällen nicht gerechtfertigt ist und zu verzerrten Parameterschätzungen führen kann. Im folgenden sollen Modelle vorgestellt werden, die von unterschiedlichen Regressionskonstanten ausgehen. Zur Begründung dieser Annahme sei noch einmal das Beispiel der Wahlfunktion betrachtet, deren zu schätzende Regressionsgleichung lautet:

$$SPD_{it} = \alpha + \beta \cdot ALQ_{it} + e_{it}$$

In der Regressionsanalyse wird angenommen, daß die Residuen e_{it} den Einfluß von sogenannten latenten bzw. impliziten Variablen wiedergeben. Diese Variablen üben zwar einen Einfluß auf die zu erklärende Variable aus, werden aber nicht in die Regressionsfunktion aufgenommen, da man ihren Einflüssen nur geringe Bedeutung zuschreibt oder aber weil diese Variablen nicht meßbar bzw. nicht beobachtbar sind. Die Residuen als Summe dieser Einflüsse werden als zufällig interpretiert, und ihre einzelnen Ausprägungen "springen" zwischen den einzelnen Beobachtungspunkten in einem nicht zu beobachteten Wertebereich hin und her. Diese Variation wird bei einem Panel-Datensatz sowohl unabhängig von dem Individuum als auch der Zeit angenommen. Allerdings könnten die Residuen auch den Einfluß latenter Variablen beinhalten, die zwar von Individuum zu Individuum verschieden, aber zeitkonstant sind, oder umgekehrt für jedes Individuum gleich, aber von Zeitpunkt zu Zeitpunkt verschieden.

Im Rahmen der Schätzung obiger Wahlfunktion wäre das mit der geographischen Lage eines Wahlkreises verbundene "traditionelle Wählerverhalten" ein Beispiel für eine latente

Variable W_i, die offensichtlich individuen-spezifische, aber zeitkonstante Ausprägungen annimmt. Umgekehrt wäre die "politische Stimmung" ein Beispiel für eine latente Variable Z_t, die zu verschiedenen Zeitpunkten divergierende, jedoch für alle Individuen gleiche Werte annimmt. Bei Berücksichtigung dieser latenten Variablen in den Residuen läßt sich dieser schreiben als

$$e_{it} = \mu W_i + \lambda Z_t + v_{it}$$

Allerdings liegen für die latenten Variablen keine Beobachtungen vor und somit ist eine Schätzung ihrer Parameter μ und λ nicht möglich. (Natürlich könnte man versuchen, die latenten Variablen explizit zu modellieren, indem man etwa den Anteil an Wählern mit einer bestimmten Konfessionszugehörigkeit eines Wahlkreises, die im Zeitablauf (nahezu) konstant sein, sich aber zwischen den Wahlkreisen unterscheiden dürfte, als Indikator für das "traditionelle Wählerverhalten" oder aber die Konjunkturlage als Ausdruck der "politischen Stimmung" bestimmt, aber es sei angenommen, daß letztendlich auch mit diesen Variablen die Einflüsse nur unzureichend quantifiziert und gemessen werden können.) Faßt man die Produkte $\mu_i = \mu W_i$ bzw. $\lambda_t = \lambda Z_t$ nun zu sogenannten Individual- und Zeiteffekten zusammen und ordnet die zu schätzende Regressionsfunktion um,

$$SPD_{it} = \alpha + \beta ALQ_{it} + \mu W_i + \lambda Z_t + v_{it} = \alpha + \mu_i + \lambda_t + \beta ALQ_{it} + v_{it} = \alpha_{it} + \beta ALQ_{it} + v_{it}$$

so erhält man ein Modell mit unterschiedlichen Regressionskonstanten. Je nachdem, ob man die Individual- und Zeiteffekte als feste Größen, und somit die unterschiedlichen Regressionskonstanten als schätzbare Parameter ansieht, oder aber als Zufallsvariablen betrachtet, ergibt sich eine weitere Unterteilung in Fixed Effects- und Random Effects-Modelle, die im folgenden näher vorgestellt werden sollen. Dabei soll vereinfachend zuerst einmal nur von dem Vorliegen von Individualeffekten μ_i ausgegangen werden.

2. Das "Least Squares Dummy Variable"-Modell

a) Einführung von Dummy-Variablen

Eine in der Regressionsanalyse übliche Methode zur Erfassung von unterschiedlichen Regressionskonstanten besteht in der Einführung von sogenannten Dummyvariablen, die jeweils den Wert 1 bei Vorliegen einer bestimmten qualitativen Ausprägung und sonst den Wert 0 annehmen. In unserem Fall ist die qualitative Ausprägung jeweils durch das einzelne Individuum gegeben, und somit sind *N* Dummy Variablen mit der folgenden Definition einzuführen

$$D_{jt} = \begin{cases} 1 & \textbf{falls } j = i \\ 0 & \textbf{falls } j \neq i \end{cases}$$

Die zu schätzende Regressionsgleichung nimmt folgende Gestalt an

$$y_{it} = \sum_{j=1}^{N} \beta_{1j} D_{jt} + \sum_{k=2}^{K} \beta_k x_{k,it} + v_{it}$$

Gilt für das betrachtete *i*te Individuum *i=j*, so nimmt die Dummy-Variable zu jedem Zeitpunkt den Wert 1 an, während alle anderen Dummy-Variablen den Wert 0 aufweisen, und der geschätzte Koeffizient β_{1j} gibt die Regressionskonstante des betreffenden Individuums an. Um dies zu verdeutlichen, sei das entstehende Gleichungssystem wiedergegeben:

$$\begin{aligned}
y_{11} &= \beta_{11}1+\beta_{12}0+\ldots+\beta_{1N}0+\beta_2 x_{2,11}+\beta_3 x_{3,11}+\ldots+\beta_K x_{K,11}+e_{11}\\
y_{12} &= \beta_{11}1+\beta_{12}0+\ldots+\beta_{1N}0+\beta_2 x_{2,12}+\beta_3 x_{3,12}+\ldots+\beta_K x_{K,11}+e_{12}\\
&\vdots\\
y_{1T} &= \beta_{11}1+\beta_{12}0+\ldots+\beta_{1N}0+\beta_2 x_{2,1T}+\beta_3 x_{3,1T}+\ldots+\beta_K x_{K,1T}+e_{1T}\\
y_{21} &= \beta_{11}0+\beta_{12}1+\ldots+\beta_{1N}0+\beta_2 x_{2,21}+\beta_3 x_{3,21}+\ldots+\beta_K x_{K,21}+e_{21}\\
y_{22} &= \beta_{11}0+\beta_{12}1+\ldots+\beta_{1N}0+\beta_2 x_{2,22}+\beta_3 x_{3,22}+\ldots+\beta_K x_{K,22}+e_{22}\\
&\vdots\\
y_{2T} &= \beta_{11}0+\beta_{12}1+\ldots+\beta_{1N}0+\beta_2 x_{2,2T}+\beta_3 x_{3,2T}+\ldots+\beta_K x_{K,2T}+e_{2T}\\
&\vdots\\
y_{N1} &= \beta_{11}0+\beta_{12}0+\ldots+\beta_{1N}1+\beta_2 x_{2,N1}+\beta_3 x_{3,N1}+\ldots+\beta_K x_{K,N1}+e_{N1}\\
y_{N2} &= \beta_{11}0+\beta_{12}0+\ldots+\beta_{1N}1+\beta_2 x_{2,N2}+\beta_3 x_{3,N2}+\ldots+\beta_K x_{K,N2}+e_{N2}\\
&\vdots\\
y_{NT} &= \beta_{11}0+\beta_{12}0+\ldots+\beta_{1N}1+\beta_2 x_{2,NT}+\beta_3 x_{3,NT}+\ldots+\beta_K x_{K,NT}+e_{NT}
\end{aligned}$$

Nimmt man für die Residuen die klassischen Annahmen als gegeben an, so kann dieses Modell wie ein gewöhnlicher Gleichungssatz mit Hilfe des OLS-Verfahrens geschätzt werden. Die sich ergebenden Parameterwerte für die Dummy-Variablen können wie alle anderen Parameter behandelt werden, sind also den üblichen Testmethoden zugänglich.

b) Bildung von Durchschnitten

Die Schätzung des Modells bei Einführung von *N* Dummy-Variablen ist mit der Inversion einer Matrix der Ordnung (*N+K-1*) verbunden, was insbesondere bei großem *N*, also einer Stichprobe mit vielen Untersuchungseinheiten, zu numerischen Problemen führen kann. Es läßt sich jedoch eine andere Form der Darstellung finden, die diesen Nachteil vermei-det, indem der Vektor der Regressionsgewichte getrennt von dem der Regressionskonstante geschätzt wird. Dabei macht man sich die spezielle,sich nach Einführung von Dummy-Variablen ergebende Struktur der Matrix der erklärenden Variablen zunutze, indem man die

Formel für die sogenannte partionierte OLS-Schätzung verwendet. Zur Veranschaulichung sei noch einmal von der Gleichung ausgegangen

$$y_{it} = \sum_{j=1}^{N} \beta_{1j} D_{jt} + \sum_{k=2}^{K} \beta_k x_{k,it} + v_{it}$$

Bildet man für jedes Individuum einen zeitlichen Durchschnitt der Beobachtungswerte, so erhält man folgenden Ausdruck

$$\overline{y}_{i.} = \sum_{j=1}^{N} \beta_{1j} D_{jt} + \sum_{k=2}^{K} \beta_k \overline{x}_{k,i.} + \overline{v}_{i.}$$

mit

$$\overline{y}_{i.} = \frac{1}{T}\sum_{t=1}^{T} y_{it} \quad \overline{x}_{k,i.} = \frac{1}{T}\sum_{t=1}^{T} x_{k,it} \quad \overline{v}_{i.} = \frac{1}{T}\sum_{t=1}^{T} v_{it}$$

Da D_{jt} für (i=j) zu jedem Zeitpunkt den Wert 1 annimmt, gilt $D_{jt} = \overline{D}_{j.}$. Subtrahiert man nun die beiden Gleichungen voneinander, so ergibt sich

$$y_{it} - \overline{y}_{i.} = \sum_{k=2}^{K} \beta_k (x_{k,it} - \overline{x}_{k,i.}) + (v_{it} - \overline{v}_{i.})$$

Wendet man auf diese Gleichung eine OLS-Schätzung ohne Regressionskonstante an, so ist die Schätzung des Parametervektors $\mathbf{ß} = (\beta_2\ \beta_3 \ldots \beta_K)'$ identisch mit der Schätzung der Regressionsgewichte, die sich bei Einführung von Dummy-Variablen ergeben hätte. Diese Schätzung der Regressionsgewichte, die die interindividuelle Streuung durch Bildung der Durchschnitte eliminiert und nur die Streuung innerhalb eines Individuums berücksichtigt, wird als "within"-Schätzung bezeichnet. Jede der Variablen wird hier als Abweichung von individuell verschiedenen Mittelwerten ausgedrückt. Nach der Schätzung der Regressionsgewichte können die individuellen Regressionskonstanten über folgende Formel geschätzt werden.

$$\beta_{1i} = \overline{y}_{i.} - \beta_2 \overline{x}_{2,i.} - \beta_3 \overline{x}_{3,i.} - \ldots - \beta_K \overline{x}_{K,i.} = \overline{y}_{i.} - \sum_{k=2}^{K} \beta_k \overline{x}_{k,i.}$$

Eine Schätzung der Varianz der Residuen kann über die residuale Abweichungsquadratsumme vorgenommen werden, die sich unabhängig von dem gewählten Schätzansatz (Einführung von Dummy-Variablen oder Bildung von Durchschnitten) ergibt.

$$\hat{\sigma}_v^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{v}_{it}^2}{NT - (N + K - 1)}$$

Im Ergebnis bleibt festzuhalten, daß im "LSDV"-Modell über die Einführung von Dummy-Variablen Individualeffekte berücksichtigt werden können. Ist die Zahl der Individuen groß, empfiehlt sich eine getrennte Schätzung von Regressionskonstanten und Regressionsgewichten. Dies kann über eine einfache Transformation der Beobachtungswerte erreicht werden, bei der jede Variable als Abweichung von individuell verschiedenene Mittelwerten ausgedrückt wird. Da so jeweils nur die zeitliche Streuung innerhalb der Individuen berücksichtigt wird, bezeichnet man diesen Schätzer auch als "within"-Schätzer.[2]

c) Test auf Vorliegen von Individualeffekten

Die naheliegende Frage, ob überhaupt von dem Vorliegen von Individualeffekten bzw. von unterschiedlichen Regressionskonstanten ausgegangen werden kann, läßt sich mit Hilfe eines F-Tests entscheiden, der die nichterklärten Abweichungsquadratsummen der Residuen miteinander vergleicht, die sich jeweils bei Verwendung des "Classical-Pooling"-Modells bzw. des "LSDV"-Modells ergeben. Die Nullhypothese lautet dementsprechend

$$H_0 = \beta_{11} = \beta_{12} = \ldots = \beta_{1N}$$

während die Alternativhypothese sich als

H_1: nicht alle der β_{1i} sind gleich

formulieren läßt. Die zugehörige Teststatistik

$$F = \frac{\dfrac{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}^2 - \sum_{i=1}^{N}\sum_{t=1}^{T}\hat{v}_{it}^2}{(N-1)}}{\dfrac{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{v}_{it}^2}{(NT-(N+K-1))}}$$

ist bei Gültigkeit der Nullhypothese F-verteilt mit $(N-1,\ NT-(N+K-1))$ Freiheitsgraden, wobei $\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}^2$ die Abweichungsquadratsumme der Residuen des "Classical Pooling" -Modells ist. Dieses wird auch als das restringierte Modell bezeichnet, da hier die Restriktion gleicher Regressionskonstanten für alle Individuen vorausgesetzt wird. Die Freiheitsgrade im Zähler (*N* - 1) entsprechen dabei der Zahl der eingeführten Restriktionen. Die Abweichungsquadratsumme des unrestringierten Modells wird durch die Quadratsumme der

[2] Voraussetzung ist hierbei natürlich, daß eine zeitliche Streuung gegeben ist, die nicht durch Meßfehler hervorgerufen wird.

Residuen des "LSDV"-Modells $\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{v}_{it}^2$ gegeben. Die Freiheitsgrade im Nenner $(NT-(N+K-1))$ entsprechen den Freiheitsgraden des unrestringierten Modells (Stichprobengröße minus der Zahl der zu schätzenden Parameter).

Übersteigt der empirische F-Wert bei einem vorgegebenen Signifikanzniveau $(1-\alpha)$ den kritischen F-Wert, so ist die Nullhypothese gleicher Regressionskonstanten abzulehnen und dementsprechend das "LSDV"-Modell vorzuziehen. ***Judge*** et al. (1988, S.476) weisen darauf hin, die Signifikanz der Dummyvariablen nicht über die jeweiligen t-Werte, sondern mit dem obigen F-Test, der dementsprechend auch als mehrfach partieller F-Test bezeichnet wird, zu überprüfen, da sich sonst unter bestimmten Umständen unterschiedliche Empfehlungen bezüglich der Signifikanz der individuellen Dummy-Variablen ergeben.

2. "Error-Components"-Modell

a) Herleitung der GLS-Schätzung

Bei Verwendung des "LSDV"-Modells konnte für jedes Individuum eine eigene Regressionskonstante geschätzt werden, weil die Individualeffekte als feste Größen interpretiert wurden.

$$y_{it}=\beta_1+\mu_i+\sum_{k=2}^{K}\beta_k x_{k,it}+v_{it}=\beta_{1i}+\sum_{k=2}^{K}\beta_k x_{k,it}+v_{it}$$

Im "Error Components"-Modell ("EC"-Modell) werden demgegenüber die Individualeffekte als zufällig betrachtet, so daß auch die individuellen Regressionskonstanten $\beta_{1i}=\overline{\beta}_1+\mu_i$ als Zufallsvariable zu interpretieren sind. Über die zufälligen Individualeffekte werden folgende Annahmen gemacht

$$E(\mu_i)=0 \quad E(\mu_i^2)=\sigma_\mu^2 \quad E(\mu_i\mu_j)=0$$

und weiterhin, daß die zufälligen Individualeffekte mit den Residuen unkorreliert sind. Die Individuen werden als eine Stichprobe aus einer größeren Grundgesamtheit betrachtet und das Ziel ist nun eine Schätzung des für diese Grundgesamtheit gültigen Parametervektors $\boldsymbol{ß}=(\overline{\beta}_1\ \beta_2\ldots\beta_K)'$.

Die Schätzgleichung lautet

$$y_{it}=\overline{\beta}_1+\sum_{k=2}^{K}\beta_k x_{k,it}+\mu_i+v_{it}=\overline{\beta}_1+\sum_{k=2}^{K}\beta_k x_{k,it}+w_{it}$$

wobei sich nun für den Mittelwert und die Varianzen bzw. Kovarianzen der Residuen w_{it} folgende Ausdrücke ergeben

$$E(w_{it}) = E(\mu_i) + E(v_{it}) = 0$$

$$E(w_{it}^2) = E(\mu_i^2) + E(v_{it}^2) + 2E(\mu_i v_{it}) = \sigma_\mu^2 + \sigma_v^2$$

wegen $E(\mu_i v_{it}) = 0$

$$E(w_{it} w_{is}) = E(\mu_i^2) + E(v_{it} v_{is}) + E(\mu_i v_{it}) + E(\mu_i v_{is}) = \sigma_\mu^2$$

wegen $E(v_{it} v_{is}) = 0$ für $t \neq s$

Ferner ist $E(w_{it} w_{js}) = E(\mu_i \mu_j) + E(v_{it} v_{is}) + E(\mu_j v_{it}) + E(\mu_i v_{js}) = 0$ für $t \neq s, i \neq j$

Die Varianz-Kovarianzmatrix der Residuen besitzt für jedes einzelne Individuum die gleiche Darstellung

$$\mathbf{V_i} = \begin{bmatrix} \sigma_\mu^2 + \sigma_v^2 & \sigma_\mu^2 & \cdots & \sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_\mu^2 + \sigma_v^2 & \cdots & \sigma_\mu^2 \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_\mu^2 & \sigma_\mu^2 & \cdots & \sigma_\mu^2 + \sigma_v^2 \end{bmatrix}_{T \times T}$$

aus der hervorgeht, daß sie die Annahme der Homoskedastizität (konstante Varianzen) erfüllen, allerdings über verschiedene Beobachtungszeitpunkte miteinander korreliert sind. Im Gegensatz zu dem von ***Kmenta*** betrachteten Fall eines autoregressiven Prozesses bleibt die Größenordnung dieser Korrelation jedoch über die Zeit konstant. Die Varianz-Kovarianz-matrix der Residuen für alle Individuen ist dementsprechend zwar blockdiagonal mit den über $\mathbf{V_i}$ gegebenen Blöcken, allerdings besitzt sie nicht die Darstellung einer den klassischen Annahmen entsprechenden Diagonalmatrix $\sigma_e^2 \mathbf{I}_{NT}$.

Deswegen ist für die Schätzung des Parametervektors das schon bei den ***Kmenta***-Modellen gezeigte GLS-bzw. EGLS-Verfahren anzuwenden. Auch hier besteht die Aufgabe in der Formulierung einer geeigneten Transformationsmatrix **P** und der Schätzung ihrer Elemente. Es läßt sich nun zeigen, daß tatsächlich die Varianz-Kovarianzmatrix der Residuen w_{it} für das "EC"-Modell eine Darstellung besitzt, die zu einer Transformation der Beobachtungswerte gemäß

$$\mathbf{Py} = \mathbf{PX}ß + \mathbf{Pe} \text{ bzw. } \mathbf{y}^* = \mathbf{X}^* ß + \mathbf{e}^*$$

führt, so daß eine anschließende OLS-Schätzung auf **y*** und **X*** der GLS-Schätzung entspricht. Ohne auf die Herleitung dieser Matrix einzugehen, seien nur die notwendigen Transformationen angeführt

$y_{it}^{ec} = y_{it} - \theta \bar{y}_{i.}$ und $x_{k,it}^{ec} = x_{k,it} - \theta \bar{x}_{k,i.}$

wobei $\theta = 1 - \frac{\sigma_v}{\sigma_1}$ mit $\sigma_1 = \sqrt{T\sigma_\mu^2 + \sigma_v^2}$

Dementsprechend ergibt sich der GLS-Schätzer über eine OLS-Schätzung folgender Gleichung

$$y_{it} - \theta \bar{y}_{i.} = (1-\theta)\bar{\beta}_1 + \sum_{k=2}^{K} \beta_k (x_{k,it} - \theta \bar{x}_{k,i.}) + w_{it}$$

Stellt man diese Gleichung der Schätzgleichung des "within"-Schätzers

$$y_{it} - \bar{y}_{i.} = \sum_{k=2}^{K} \beta_k (x_{k,it} - \bar{x}_{k,i.}) + (v_{it} - \bar{v}_{i.})$$

gegenüber, so wird die Beziehung zwischen diesen beiden Ansätzen und die Rolle der einzelnen Komponenten der Residuen ("Error Components") über den Faktor θ deutlich. Auch beim "EC"-Modell sind die transformierten Beobachtungswerte als Abweichungen vom individuellen Mittelwert aufzufassen, allerdings wird dieser vorher mit einem Faktor gewichtet, der sich als das Verhältnis von zwei Varianzen ergibt. Es läßt sich zeigen, daß der Ausdruck für den GLS-Schätzer einen gewichteten Durchschnitt aus zwei OLS-Schätzern darstellt, wobei sich die Gewichte aus der Größenordnung der jeweiligen Varianzkomponenten ergeben. Dabei ist der eine OLS-Schätzer der bekannte "within" Schätzer, der die Informationen über die Streuung innerhalb der Individuen ausnutzt, während der andere OLS-Schätzer als "between"-Schätzer bezeichnet wird, der die Streuung zwischen den Individuen ausnutzt. Dieser ergibt sich aus einer OLS-Schätzung über die *N* individuellen Mittelwerte

$$\bar{\mathbf{y}} = \bar{\mathbf{X}}\boldsymbol{ß} + \bar{\mathbf{w}}$$

wobei $\bar{\mathbf{y}}, \bar{\mathbf{X}}$ und $\bar{\mathbf{w}}$ jeweils die typischen Elemente $\bar{y}_{i.}, \bar{x}_{k,i.}$ und $\bar{w}_{i.}$ enthalten.

Der Faktor θ entscheidet nun über den Einfluß dieser Streuungsarten, der bei der GLS-Schätzung berücksichtigt wird. Aus der Definition ergibt sich, daß θ bei großem *T* oder großem σ_μ relativ zu σ_v gegen den Wert 1 strebt. Dies ist gleichbedeutend mit einem großen Einfluß des "within"-Schätzers, im Grenzfall *T* gegen ∞ stimmen die Schätzwerte für **ß** im "LSDV"-Modell mit denen des "EC"-Modells überein. Umgekehrt, bei relativ großem σ_v gegenüber σ_μ, tendiert θ gegen 0 und der GLS-Schätzer stimmt mit dem OLS-Schätzer des "Classical Pooling"-Modells überein.

b) Die EGLS-Schätzung

In der Praxis jedoch sind die Varianzen der Residuenkomponenten nicht bekannt, so daß man vom GLS- zum EGLS-Verfahren übergeht, in dem eine Schätzung der Varianz-Kovarianzmatrix **V** bzw. der Transformationsmatrix **P** vorgenommen wird. Dies ist gleichbedeutend mit einer Schätzung der Varianzkomponenten bzw. ihrer Quadratwurzeln σ_v und σ_μ zur Bildung der gesuchten Größe θ und anschließender Transformation der Beobachtungswerte, wie oben gezeigt. Dabei macht man sich nun die schon oben erwähnten "within"- und "between"- Schätzer zunutze:

Es kann gezeigt werden, daß die beim "LSDV"-Modell aus den Residuen erhaltene ("within") Schätzung der Varianz $\hat{\sigma}_v^2$ der Residuen eine unverzerrte Schätzung für σ_v^2 darstellt. Um eine Schätzung für den Term σ_μ^2 zu erhalten , bildet man den "between"-Schätzer, der einer OLS-Schätzung über die *N* individuellen, über die Zeit gemittelten Werte der Beobachtungsvariablen entspricht.

$$\overline{y}_{i.} = \overline{\beta}_1 + \sum_{k=2}^{K} \beta_k \overline{x}_{k,i.} + \overline{w}_{i.}$$

Bildet man die Varianz der Residuen $\overline{w}_{i.} = \mu_i + \overline{v}_{i.}$ so ergibt sich aufgrund der postulierten Unabhängigkeitsannahme

$$\text{var}(\mu_i + \overline{v}_{i.}) = \sigma_\mu^2 + \frac{\sigma_v^2}{T} = \frac{\sigma_1^2}{T}$$

Zur Schätzung der Varianz können die sich bei Durchführung der "between"-Schätzung ergebenden Residuen verwendet werden. Der so erhaltene Varianzschätzer mit *T* multipliziert, ergibt dann den gesuchten Ausdruck für die zweite Varianzkomponente in θ. Transformiert man die Beobachtungswerte von **y** und **X** gemäß

$$y_{it}^{ec} = y_{it} - \hat{\theta}\overline{y}_{i.} \text{ und } x_{k,it}^{ec} = x_{k,it} - \hat{\theta}\overline{x}_{k,i.}$$

$$\text{wobei } \hat{\theta} = 1 - \frac{\hat{\sigma}_v}{\hat{\sigma}_1} \text{ mit } \hat{\sigma}_1 = \sqrt{T\hat{\sigma}_\mu^2 + \hat{\sigma}_v^2}$$

und führt mit den transformierten Variablen eine OLS-Schätzung durch, erhält man den gesuchten EGLS-Schätzer. Dabei ist allerdings zu beachten, daß auch die Regressionskonstante entsprechend transformiert wird. Anstelle der Einsen in der Matrix der erklärenden Variablen sind also die Werte $1-\hat{\theta}$ einzugeben. Die Vorgehensweise zur Ermittlung des EGLS-Schätzers des "EC"-Modells läßt sich zusammengefaßt darstellen:

- Berechnung des "LSDV"-Modells, entweder über Einführung von Dummy-Variablen oder über die transformierten Beobachtungswerte, ausgedrückt als Abweichung vom in-

dividuellen Mittelwert, wobei letztere Variante rechentechnische Vorteile bezüglich der später durchzuführenden Transformationen besitzt. Ermittlung der Quadratsumme der Residuen des "LSDV"-Modells, die bei beiden Vorgehensweisen identisch ist, und Schätzung der Varianzkomponente $\hat{\sigma}_v^2$ über

$$\hat{\sigma}_v^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{v}_{it}^2}{(NT-(N+K-1))}$$

- Berechnung des "between"-Schätzers über eine OLS-Regression bei Verwendung der *N* individuellen Mittelwerte von erklärter und erklärenden Variablen zur Ermittlung der Quadratsumme der Residuen dieses Modells. Anschließende Schätzung der Varianzkomponente $\hat{\sigma}_1^2 = T\hat{\sigma}_\mu^2 + \hat{\sigma}_v^2$ mit Hilfe dieser Residuen

$$\frac{\hat{\sigma}_1^2}{T} = \frac{\sum_{i=1}^{N}\hat{\bar{w}}_{i.}^2}{N-K}$$

- Berechnung von $\hat{\theta} = 1 - \frac{\hat{\sigma}_v}{\hat{\sigma}_1}$

- Bildung der transformierten Beobachtungswerte

$y_{it}^{ec} = y_{it} - \hat{\theta}\bar{y}_{i.}$ und $x_{k,it}^{ec} = x_{k,it} - \hat{\theta}\bar{x}_{k,i.}$

- Eine OLS-Regression über die entsprechend transformierten Beobachtungswerte ist gleichbedeutend mit der gesuchten EGLS-Schätzung

c) Test auf Vorliegen von Individualeffekten

Wie beim "LSDV"-Modell kann auch für das "EC"-Modell die Frage, ob Individualeffekte vorliegen, mit Hilfe des F-Tests entschieden werden, der die Quadratsummen der Residuen des restringierten Modells ("Classical Pooling") mit denen des unrestringierten Modells ("LSDV") vergleicht. Die Bildung des "LSDV"-Modells zur Erfassung von Individualeffekten reicht für diese Entscheidung, unabhängig davon, ob man die Individualeffekte als fest oder zufällig betrachtet. Da zur praktischen Ermittlung des EGLS-Schätzers die "within"-Schätzung, also das "LSDV"-Modell, ohnehin benötigt wird, bedeutet dies keinen zusätzlichen Rechenaufwand.

Eine andere Teststatistik, die von ***Breusch*** und ***Pagan*** vorgeschlagen wurde und als ein Lagrange-Multiplikator-Test anzusehen ist, benötigt demgegenüber nur die Residuen des restringierten Modells. Die Nullhypothese lautet

$$H_0: \quad \sigma_\mu^2 = 0$$

und unter ihr ist die Teststatistik

$$LM = \frac{NT}{2(T-1)} \left(\frac{\sum_{i=1}^{N} \left(\sum_{t=1}^{T} \hat{e}_{it} \right)^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} \hat{e}_{it}^2} - 1 \right)^2$$

asymptotisch $\chi^2_{(1)}$ -verteilt. Dabei sind die $\hat{e}_{it}$'s die Residuen des "Classical-Pooling"-Modells. Übersteigt der empirische $\tilde{\chi}^2$-Wert bei vorgegebenem Signifikanzniveau $(1-\alpha)$ den kritischen $\chi^2_{(1)}$-Wert, ist die Nullhypothese abzulehnen, und es kann auf Vorliegen von Individualeffekten geschlossen werden.

4. Fixed versus Random Effects

Eine offensichtliche Frage ist, ob bei Vorliegen von Individualeffekten diese als fest oder als zufällig anzusehen sind, ob also das "LSDV"-Modell oder das "EC"-Modell Anwendung finden sollte. Ein erster Gesichtspunkt betrifft die Größenordnung der Stichprobe. Wenn *T* bei kleinem *N* sehr groß wird, so gehen die beiden Modelle ineinander über, und das rechentechnisch einfachere "LSDV"-Modell dürfte vorzuziehen sein. Trifft dieser Fall jedoch nicht zu, so können erhebliche Unterschiede zwischen den Modellen erwartet werden und andere Aspekte müssen bei der Entscheidung berücksichtigt werden.

Bei der Herleitung dieser Modelle wurden Individualeffekte als Komponente der ursprünglichen Residuen aufgefaßt, die den Einfluß latenter Variablen auf die zu erklärende Variable wiedergeben. Während man den Einfluß der latenten Variablen, die nicht nur über die Zeit, sondern auch über die Individuen variieren, weiterhin im "LSDV"-Modell in einem als zufällige Größe zu interpretierenden Residuen v_{it} erfaßt, werden die zeitinvarianten und für jedes Individuum unterschiedlichen latenten Variablen als feste Individualeffekte angesehen, deren Einfluß über die Parameter der Dummy-Variablen geschätzt wird. Allerdings liefern diese Dummy-Variablen keine Informationen über die Ursachen, welche zu einer Niveauverschiebung der Regressionsfunktion für jedes Individuum führen, ihre Parameter messen lediglich deren Einfluß auf Kosten eines Verlustes an Freiheitsgraden. Insofern scheint es angebrachter, die bestehende Unkenntnis über die Individualeffekte ähnlich der über die anderen latenten Variablen zu behandeln, das heißt, diese als ebenfalls zufällig zu betrachten.

In der ökonometrischen Literatur wird jedoch darauf hingewiesen, daß die Individualeffekte immer als zufällige Größen interpretiert werden können, da sie erst nach der Ziehung der

Stichprobe bekannt sind bzw. geschätzt werden können. Es ändert sich lediglich die Sichtweise, mit der Schlußfolgerungen über diese Stichprobe gezogen werden können: Bei dem Fixed Effects-Modell ist die statistische Inferenz bedingt (abhängig) von den jeweiligen Individuen in der Stichprobe, während bei der Annahme von Random Effects eine Inferenz unbedingt (unabhängig) von den Individuen dieser Stichprobe erfolgt. Damit werden Aussagen über die dahinterstehende Grundgesamtheit ermöglicht. Der Anwender muß folglich über die Art der zu treffenden Aussagen entscheiden. Bedingte Inferenz sollte dann getroffen werden, wenn die Individuen nicht als eine Stichprobe aus einer übergeordneten Grundgesamtheit zu betrachten sind oder wenn es insbesondere die in der Stichprobe enthaltenen Individuen sind, über die er Aussagen treffen will. Sollen demgegenüber die Schlußfolgerungen die Grundgesamtheit betreffen und können die Individuen als eine Stichprobe hieraus angesehen werden, dann empfiehlt sich die unbedingte Inferenz mit Hilfe der Annahme von Random Effects.

Im Gegensatz zum Fixed Effects-Modell müssen hierzu jedoch restriktive Annahmen über die Verteilung der Zufallsvariablen μ gemacht werden, so daß nur bei deren Gültigkeit diese zusätzliche Information zu einem Effizienzgewinn des "EC"-Modells führen kann.

Deshalb sollte die Eignung der Annahme überprüft werden, daß die Individualeffekte identisch und unabhängig verteilte Zufallsvariablen mit einem Mittelwert von Null und konstanter Varianz sind. So kann zum Beispiel gezeigt werden, daß bei einer Korrelation zwischen den erklärenden Variablen und den Individualeffekten der EGLS-Schätzer des "EC"-Modells verzerrt und inkonsistent ist, während gerade in dieser Situation sich der "LSDV"-Schätzer als effizient erweist.

So könnte man sich im Rahmen der Schätzung einer Wahlfunktion zum Beispiel vorstellen, daß die mit einer bestimmten geographischen Lage verbundenen Individualeffekte eines Wahlkreises in Form des "traditionellen Wählerverhaltens" über die vorherrschende Wirtschaftsstruktur in diesem Gebiet mit der Arbeitslosigkeit korreliert sind, da in einem landwirtschaftlich strukturierten Gebiet einerseits eher konservativ gewählt wird, während die Arbeitslosigkeit tendenziell geringer als in industriellen Ballungszentren ausfällt.

Um die Hypothese einer bestehenden Korrelation zwischen erklärenden Variablen und Individualeffekten zu überprüfen, kann ein von ***Hausman*** entwickeltes Testverfahren angewandt werden, wobei man sich die unterschiedlichen Eigenschaften des "LSDV"- und des "EC"-Schätzers zunutze macht.Unter der Nullhypothese fehlender Korrelation ist der"EC"-Schätzer unverzerrt und effizient hingegen bei Vorliegen einer Korrelation verzerrt, während der "LSDV"-Schätzer sowohl bei Gültigkeit der Nullhypothese wie auch bei bestehender Korrelation konsistent ist. Deswegen kann bei Gültigkeit der Nullhypothese erwartet werden, daß, zumindest asymptotisch, die beiden Schätzer nur zufällig voneinander abweichen werden, während diese Abweichung bei bestehender Korrelation weitaus größer

sein dürfte. Im wesentlichen wird beim ***Hausman***-Test also der sich empirisch ergebende Unterschied beider Schätzer auf Signifikanz geprüft.

Zur Durchführung des ***Hausman***-Tests eignet sich besonders die F-Test Version, bei der wiederum die Quadratsummen von zwei Modellen miteinander verglichen werden. Das eine Modell ist dabei das "EC"-Modell, während das andere eine Kombination des "EC"- und des "LSDV"-Modells darstellt,

$$y_{it} - \theta\,\bar{y}_{i.} = (1-\theta)\bar{\beta}_1 + \sum_{k=2}^{K} \beta_k^{EC}(x_{k,it} - \theta\,\bar{x}_{k,i.}) + \sum_{k=2}^{K} \beta_k^{LSDV}(x_{k,it} - \bar{x}_{k,i.}) + w^*_{it}$$

indem man zusätzlich zu den bereits im Ansatz befindlichen Regressoren des "EC"-Modells die des "LSDV"-Modells einbezieht.

Die Nullhypothese lautet

$$H_0 : \beta_k^{LSDV} = 0 \quad \text{gegen} \quad H_1 : \beta_k^{LSDV} \neq 0$$

Man prüft nun die Signifikanz der zusätzlich eingeführten Parameter des "LSDV"-Modells über die Teststatistik

$$F = \frac{\dfrac{\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{w}_{it}^2 - \sum_{i=1}^{N}\sum_{t=1}^{T} \hat{w}_{it}^{2}\,*}{(K-1)}}{\dfrac{\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{w}_{it}^{2}\,*}{(NT-2K+1)}}$$

dabei entspricht $\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{w}_{it}^2$ der Quadratsumme der Residuen des restringierten Modells, also des "EC"-Modells, mit $(K-1)$ als der Zahl der eingeführten Restriktionen, während $\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{w}_{it}^{2}\,*$ die Quadratsumme der Residuen des unrestringierten Modells, also des kombinierten "EC"- und "LSDV"-Modells, und $(NT-2K+1)$ die Zahl der Freiheitsgrade des unrestringierten Modells angeben.

Unter der Nullhypothese ist die Teststatistik Fverteilt mit den angegebenen Freiheitsgraden von Zähler und Nenner. Übersteigt der empirische F-Wert den bei einem vorgegebenen Signifikanzniveau ermittelten kritischen F-Wert, so ist die Nullhypothese abzulehnen. Dies bedeutet, daß man von einer Korrelation zwischen erklärenden Variablen und Individual-

effekten ausgehen kann, und somit das "LSDV"-Modell dem "EC"-Modell in dieser Situation vorzuziehen ist.

Zusammenfassend schlagen ***Judge*** et al. (1985, S.527) vor, daß "a reasonable prescription is to use the error components model if the μ_i ~ i.i.d. $(0, \sigma_\mu^2)$ assumption is a reasonable one and *N* is sufficiently large for reliable estimation of σ_μ^2; otherwise, particularly when μ_i and $\mathbf{X}_i$ are correlated, or *N* is small, use the dummy variable (within) estimator." Auf die Frage, wie groß *N* sein sollte, führen sie an anderer Stelle (1989, S.490) eine Untersuchung von ***Taylor*** (1980) an, der gezeigt hat, daß selbst bei relativ geringem Stichprobenumfang $\left[T \geq 3, N - K \geq 9\,;\, T \geq 2, N - K \geq 10\right]$ die Effizienzvorteile des "EC"-Schätzers zum Tragen kommen.

5. Berücksichtigung von Zeiteffekten

a) Zeiteffekte im "LSDV"-Modell

Bisher wurde zur Vereinfachung von möglichen Zeiteffekten λ_t abstrahiert, deren Einführung jedoch im Rahmen der obigen Modelle eine naheliegende Erweiterung darstellt.

Im "LSDV"-Modell können entweder *T* zusätzliche Dummy-Variablen zur Erfassung der Zeiteffekte mit der folgenden Definition

$$\mathrm{D}_{\mathrm{is}} = \begin{cases} 1 & \textbf{falls } \mathrm{t} = \mathrm{s} \\ 0 & \textbf{sonst} \end{cases}$$

eingeführt werden oder die Beobachtungswerte von **y** und **X** werden wie folgt transformiert

$$y_{it}^{LSDV} = y_{it} - \bar{y}_{i.} - \bar{y}_{.t} + \bar{y}_{..}$$

$$x_{k,it}^{LSDV} = x_{k,it} - \bar{x}_{k,i.} - \bar{x}_{k,.t} + \bar{x}_{k,..}$$

wobei

$$\bar{y}_{i.} = \frac{1}{T}\sum_{t=1}^{T} y_{it} \qquad \bar{y}_{.t} = \frac{1}{N}\sum_{i=1}^{N} y_{it} \qquad \bar{y}_{..} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}$$

$$\bar{x}_{k,i.} = \frac{1}{T}\sum_{t=1}^{T} x_{k,it} \qquad \bar{x}_{k,.t} = \frac{1}{N}\sum_{i=1}^{N} x_{k,it} \qquad \bar{x}_{k,..} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{k,it}$$

Dabei wird ein einzelner Beobachtungspunkt nicht mehr nur als Abweichung vom jeweiligen individuellen Durchschnitt über alle Zeitpunkte, sondern auch als Abweichung vom jeweili-

gen zeitlichen Durchschnitt über alle Individuen ausgedrückt. Führt man eine OLS-Regression mit Individual- und Zeitdummies oder mit den transformierten Beobachtungswerten durch, so erhält man den "LSDV"-Schätzer für feste Individual- und Zeiteffekte. Auch hier kann die Einführung von Zeiteffekten mit Hilfe eines F-Tests und entsprechend definierten Quadratsummen der Residuen überprüft werden.

b) Zeiteffekte im "EC"-Modell

Für die Durchführung einer Schätzung des "EC"-Modells müssen die Beobachtungswerte ebenfalls transformiert werden

$$y_{it}^{ec} = y_{it} - \theta_1 \overline{y}_{i.} - \theta_2 \overline{y}_{.t} + \theta_3 \overline{y}_{..}$$

$$x_{k,it}* = x_{k,it} - \theta_1 \overline{x}_{k,i.} - \theta_2 \overline{x}_{k,.t} + \theta_3 \overline{x}_{k,..}$$

Dabei drücken die Koeffizienten θ wieder den Einfluß der Varianzkomponenten aus

$$\theta_1 = 1 - \frac{\sigma_v}{\sigma_1}, \quad \theta_2 = 1 - \frac{\sigma_v}{\sigma_2}, \quad \theta_3 = \theta_1 + \theta_2 - 1 + \frac{\sigma_v}{\sigma_3}$$

mit $\sigma_1^2 = \sigma_v^2 + T\sigma_\mu^2$, $\sigma_2^2 = \sigma_v^2 + N\sigma_\lambda^2$

und $\sigma_3^2 = \sigma_v^2 + T\sigma_\mu^2 + N\sigma_\lambda^2$

Eine OLS-Schätzung über die Werte **y*** und **X***, wobei zur Transformation die entsprechenden Varianzkomponenten vorher geschätzt werden müssen, entspricht der EGLS-Schätzung des "EC"-Modells mit Individual- und Zeiteffekten.

Neben der Überprüfung auf Vorliegen von Individual-und Zeiteffekten durch den schon unter a) genannten F-Test, kann der von ***Breusch*** und ***Pagan*** vorgeschlagene LM-Test mit entsprechend erweiterter Teststatistik

$$LM = \frac{NT}{2}\left\{\frac{1}{T-1}\left(\frac{\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\hat{e}_{it}\right)^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}^2} - 1\right)^2 + \frac{1}{N-1}\left(\frac{\sum_{t=1}^{T}\left(\sum_{i=1}^{N}\hat{e}_{it}\right)^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}^2} - 1\right)^2\right\} \sim \chi_2^2$$

verwandt werden.

Bei den obigen Betrachtungen wurde bisher unterstellt, daß sich die Individual- und/oder Zeiteffekte allein auf die Regressionskonstante auswirkten, während die Regressionsgewichte weiterhin als konstant angenommen wurden. In der ökonometrischen Literatur werden jedoch auch Modelle behandelt, in denen Individual- und/oder Zeiteffekte für den gesamten Parametervektor zugelassen werden (vgl. Tabelle 1). Deren Erläuterung würde allerdings den Rahmen dieser Einführung sprengen, so daß auf die unten angeführte Literatur verwiesen wird. Statt dessen soll im nächsten Abschnitt ein simplifiziertes Rechenbeispiel gegeben werden, um die in den vorangegangenen Abschnitten dargestellten Modelle zu veranschaulichen. Gleichzeitig soll anhand dieses Beispiels aufgezeigt werden, daß zur praktischen Umsetzung einer Regressionsanalyse mit Panel-Daten sich die Benutzung eines Tabellenkalkulationsprogrammes (z.B. Excel) in Kombination mit einem Statistik-Programmpaket empfiehlt, welches das Einlesen der Daten aus dem Tabellenkalkulationsprogramm unterstützt (z.B. SPSS für Windows, das explizit einen Befehl zur Eingabe von Excel-Dateien beinhaltet).

IV. Rechenbeispiel: Die Schätzung einer Wahlfunktion

Angenommen man habe in 4 Wahlkreisen für 5 Wahlperioden die Werte des jeweiligen Anteils an SPD-Wählern und die jeweilige Arbeitslosenquote beobachtet, die in Tabelle 2 wiedergegeben sind:

Tabelle 2: Anteil der SPD-Wähler und Arbeitslosenquote für 5 Wahlperioden und 4 Wahlkreise

	Wahlkreis 1		Wahlkreis 2		Wahlkreis 3		Wahlkreis 4	
Wahlperiode	SPD	ALQ	SPD	ALQ	SPD	ALQ	SPD	ALQ
1	38,46	4,32	45,52	5,13	22,86	2,39	41,86	6,49
2	35,32	5,86	48,71	4,77	18,52	1,77	44,33	6,18
3	30,78	3,85	47,01	4,68	22,93	2,86	43,21	6,05
4	35,34	4,08	50,32	7,36	25,02	3,15	46,69	7,77
5	30,83	3,99	40,05	4,28	35,13	4,01	49,7	8,12
Jahres∅	34,15	4,42	46,32	5,24	24,89	2,84	45,16	6,92

Ziel sei es, zu überprüfen, ob der Anteil der SPD-Wähler von einer Zunahme der Arbeitslosigkeit beeinflußt wird, wobei angenommen wird, daß die beiden Variablen in linearer Abhängigkeit stehen: $SPD_{it} = \alpha_i + \beta \cdot ALQ_{it} + e_{it}$

Für die Regressionsgewichte β wird für alle Kreise ein gleicher Wert erwartet; über die Regressionskonstante α_i bestehen dagegen keinerlei Vorabinformationen, so daß die Möglichkeit von individuell verschiedenen Regressionskonstanten der Wahlkreise hier in Betracht zu ziehen ist.

Das "Classical Pooling"-Modell erfordert nun nichts weiter als eine Dateneingabe, in der die entsprechenden Beobachtungswerte von abhängiger und unabhängiger Variable untereinandergeschrieben werden. Die Spalte (1) in Tabelle 3 gibt den Vektor **y** für die abhängige Variable (den Anteil der SPD-Wähler) wieder, während die Spalten (2) und (3) die Matrix **X** der unabhängigen Variablen (Konstante und Arbeitslosenquote) darstellen. Anzumerken ist, daß die Eingabe des konstanten Terms, also der mit lauter Einsen besetzten Spalte (2) von den Statistik-Programmen übernommen wird.

Für das "Classical Pooling"-Modell ergibt sich folgende Regressionsgleichung, wobei RSS die Residual Sum of Squares, also die Abweichungsquadratsumme der Residuen, angibt:

$$SPD_{it} = 14,56 + 4,75 \cdot ALQ_{it} + \hat{e}_{it} \qquad RSS = 508,26$$

Wie erwähnt liegen dem "Classical Pooling"-Modell sehr restriktive Annahmen über die Residuen zugrunde, so daß man im Rahmen des "***Kmenta***"-Modells für die 4 Wahlkreise jeweils eine verschieden große Varianz und Autokorrelation der Residuen zulassen kann. Den ersten Schritt des "***Kmenta***"-Modells haben wir mit der obigen Schätzung bereits unternommen, so daß nun aus den Residuen dieser Schätzgleichung, die in Spalte (4) aufgeführt sind, die zur Transformation der Variablen erforderlichen Werte gemäß der in Abschnitt II.2 angegebenen Formeln errechnet werden müssen. Zuerst erfolgt eine Berechnung der Autokorrelationskoeffizienten $\hat{\rho}_i$, wobei sich für den ersten Haushalt der Wert wie folgt ergibt:

$$\hat{\rho}_1 = \frac{(-7,08) \cdot 3,37 + (-2,07) \cdot (-7,08) + 1,4 \cdot (-2,07) + (-2,69) \cdot 1,4}{3,37^2 + (-7,08)^2 + (-2,07)^2 + 1,4^2} = \frac{(-15,87)}{67,73} = (-0,23)$$

Die Werte für die 3 anderen Wahlkreise lauten $\hat{\rho}_2 = 0,73$, $\hat{\rho}_3 = 0,7$ und $\hat{\rho}_4 = 0,43$.

Im nächsten Schritt sind die ursprünglichen Beobachtungswerte gemäß der Formeln in II.2 zu transformieren, wobei hier nur die Berechnung für den ersten Wahlkreis exemplarisch veranschaulicht werden soll:

$y_{11}^{*} = \sqrt{1-(-0,23)^2} \cdot 38,46 = 37,43$ $\quad x_{11}^{*} = \sqrt{1-(-0,23)^2} \cdot 4,32 = 4,2$

$y_{12}^{*} = 35,32 - (-0,23) \cdot 38,46 = 44,17$ $\quad x_{12}^{*} = 5,86 - (-0,23) \cdot 4,32 = 6,85$

$y_{13}^{*} = 30,78 - (-0,23) \cdot 35,32 = 38,9$ $\quad x_{13}^{*} = 3,85 - (-0,23) \cdot 5,86 = 5,2$

$y_{14}^{*} = 35,34 - (-0,23) \cdot 30,78 = 42,42$ $\quad x_{14}^{*} = 4,08 - (-0,23) \cdot 3,85 = 4,97$

$y_{15}^{*} = 30,83 - (-0,23) \cdot 35,34 = 38,96$ $\quad x_{15}^{*} = 3,99 - (-0,23) \cdot 4,08 = 4,93$

Die Ergebnisse für die Berechnung auch der 3 anderen Wahlkreise finden sich in den Spalten (5), (6) und (7) der Tabelle 3. Zur Spalte (6) ist eine Bemerkung angebracht. Diese ersetzt den konstanten Term in Spalte (2) des vorhergehenden Modells. Auch die mit Einsen besetzte Spalte (2) ist gemäß der angegebenen Formeln zu transformieren, so daß sich die entsprechenden Werte in Spalte (6) ergeben. Eine nochmalige Schätzung des Regressionsmodells mit den transformierten Beobachtungswerten führt zu dem folgenden Ergebnis,

$$SPD_{it}^{*} = 19,22 + 3,54 \cdot ALQ_{it}^{*} + \hat{e}_{it}^{*} \qquad RSS = 212,88$$

wobei zu beachten ist, daß bei der Schätzung die üblicherweise von den Statistikprogrammen automatisch mitgeschätzte Konstante unterdrückt wird, und an ihrer Stelle der entsprechend transformierte konstante Term $x_{1,it}^{*}$ eingegeben wird (SPSS bietet unter dem Menü-Punkt "Optionen..." mit dem Befehl "Konstante in Gleichung" die Möglichkeit, die Schätzung der Regressionskonstante zu unterdrücken). Da wir im Ansatz aber noch Heteroskedastizität vermuten, muß aus dem Residualvektor $\hat{\mathbf{e}}^{*}$, der in Spalte (8) angegeben wird, noch die für jeden Wahlkreis als unterschiedlich angenommene Varianz der Residuen $\hat{\sigma}_i^2$ bestimmt werden, für den ersten Wahlkreis lautet diese:

$$\hat{\sigma}_1^{2*} = \tfrac{1}{3}(3,9^2 + (-3,75)^2 + (-3,17)^2 + 1,17^2 + (-2,15)^2) = 15,10.$$

Für die 3 anderen Wahlkreise ergeben sich die Werte $\hat{\sigma}_2^{2*} = 31,14$, $\hat{\sigma}_3^{2*} = 19,01$ und $\hat{\sigma}_4^{2*} = 5,7$.
Die entsprechend bereinigten Werte y_{it}^{**} und $x_{k,it}^{**}$ finden sich in den Spalten (9), (10) und (11) der Tabelle 3. Führt man mit diesen Werten eine Schätzung der Regressionsfunktion durch, so erhält man für das "***Kmenta***"-Modell folgendes Ergebnis (wiederum muß die Konstante vom Anwender selbst eingelesen werden):

$$SPD_{it}^{**} = 18,07 + 3,75 \cdot ALQ_{it}^{**} + \hat{e}_{it}^{**} \qquad RSS = 11,91$$

Aus dem obigen "***Kmenta***"-Modell läßt sich das Regressionsgewicht β mit einem Wert von 3,75 ablesen, für die Regressionskonstante α ergibt sich ein Wert von 18,07.

Sowohl im "Classical Pooling"- als auch im "***Kmenta***"-Modell haben wir bisher die Gültigkeit einer identischen Regressionsfunktion für die 4 Wahlkreise angenommen. Wie in Abschnitt III.1 argumentiert, könnten wir jedoch durch die Einführung von Dummy-Variablen

Tabelle 3: Beobachtungswerte des "Classical Pooling"-Modells und transformierte Be obachtungswerte des "Kmenta"-Modells

(1) y_{it}	(2) $x_{1,it}$	(3) $x_{2,it}$	(4) $\hat{e}_{it}$	(5) y_{it}^{*}	(6) $x_{1,it}^{*}$	(7) $x_{2,it}^{*}$	(8) $\hat{e}_{it}^{*}$	(9) y_{it}^{**}	(10) $x_{1,it}^{**}$	(11) $x_{2,it}^{**}$
38,46	1	4,32	3,37	37,43	0,97	4,2	3,9	9,62	0,25	1,08
35,32	1	5,86	-7,08	44,17	1,23	6,85	-3,75	11,35	0,32	1,76
30,78	1	3,85	-2,07	38,9	1,23	5,2	-3,17	10	0,32	1,34
35,34	1	4,08	1,4	42,42	1,23	4,97	1,17	10,9	0,32	1,28
30,83	1	3,99	-2,69	38,96	1,23	4,93	-2,15	10,02	0,32	1,27
45,52	1	5,13	6,59	31,11	0,68	3,51	5,6	5,58	0,12	0,63
48,71	1	4,77	11,49	15,48	0,27	1,03	6,64	2,77	0,05	0,18
47,01	1	4,68	10,21	11,45	0,27	1,2	2,01	2,05	0,05	0,22
50,32	1	7,36	0,79	16	0,27	3,94	-3,15	2,87	0,05	0,71
40,05	1	4,28	5,15	3,32	0,27	-1,09	2	0,59	0,05	-0,2
22,86	1	2,39	-3,06	16,33	0,71	1,71	-3,37	3,75	0,16	0,39
18,52	1	1,77	-4,45	2,52	0,3	0,1	-3,6	0,58	0,07	0,02
22,93	1	2,86	-5,22	9,97	0,3	1,62	-1,54	2,29	0,07	0,37
25,02	1	3,15	-4,51	8,97	0,3	1,15	-0,87	2,06	0,07	0,26
35,13	1	4,01	1,52	17,62	0,3	1,81	5,44	4,04	0,07	0,42
41,86	1	6,49	-3,54	37,79	0,9	5,86	-0,28	15,81	0,38	2,45
44,33	1	6,18	0,41	26,33	0,57	3,39	3,36	11,02	0,24	1,42
43,21	1	6,05	-0,09	24,15	0,57	3,39	1,18	10,1	0,24	1,42
46,69	1	7,77	-4,79	28,11	0,57	5,17	-1,17	11,76	0,24	2,16
49,7	1	8,12	-3,44	29,62	0,57	4,78	1,72	12,39	0,24	2

oder aber durch die Bildung von Durchschnitten und anschließender Transformation der B-eobachtungswerte unterschiedliche Regressionskonstanten α_i zulassen. Da im allgemeinen die Zahl der Individuen (hier Wahlkreise) sehr groß sein wird, empfiehlt sich aus rechen-technischen Gründen das letztere Verfahren, welches dementsprechend hier zuerst vorge-

stellt wird. Im ersten Schritt müssen für die 4 Wahlkreise jeweils die Durchschnitte des SPD-Anteils und der Arbeitslosenquote über die 5 Wahlperioden bestimmt werden. Es ergeben sich anhand von Tabelle 2 die folgenden Werte:

$\bar{y}_{1.} = (38,46 + 35,32 + 30,78 + 35,34 + 30,83)/5 = 34,15$
$\bar{y}_{2.} = 46,32$
$\bar{y}_{3.} = 24,89$
$\bar{y}_{4.} = 45,16$

$\bar{x}_{2,1.} = (4,32 + 5,86 + 3,85 + 4,08 + 3,99)/5 = 4,42$
$\bar{x}_{2,2.} = 5,24$
$\bar{x}_{2,3.} = 2,84$
$\bar{x}_{2,4.} = 6,92$

Im nächsten Schritt müssen die ursprünglichen Daten als Abweichung der jeweiligen Mittelwerte ausgedrückt werden, für den Wahlkreis 1 ergeben sich die folgenden Werte:

$y_{11}^{lsdv} = 38,46 - 34,15 = 4,31$ $\quad x_{2,11}^{lsdv} = 4,32 - 4,42 = -0,1$
$y_{12}^{lsdv} = 35,32 - 34,15 = 1,17$ $\quad x_{2,12}^{lsdv} = 5,86 - 4,42 = 1,44$
$y_{13}^{lsdv} = 30,78 - 34,15 = -3,37$ $\quad x_{2,13}^{lsdv} = 3,85 - 4,42 = -0,57$
$y_{14}^{lsdv} = 35,34 - 34,15 = 1,19$ $\quad x_{2,14}^{lsdv} = 4,08 - 4,42 = -0,34$
$y_{15}^{lsdv} = 30,83 - 34,15 = -3,32$ $\quad x_{2,15}^{lsdv} = 3,99 - 4,42 = -0,43$

Mit den so bestimmten Werten auch für die anderen 3 Wahlkreise, die in Tabelle 4 in den Spalten (1) und (2) aufgeführt und wie im "Classical-Pooling"-Modell einfach untereinandergeschrieben sind, kann nun die Regressionsfunktion bestimmt werden, wobei allerdings wiederum zu beachten ist, daß kein konstanter Term mitgeschätzt werden darf.

Es ergibt sich folgende Schätzung für das "LSDV"-Modell:

$SPD_{it}^{lsdv} = 3,14 \cdot ALQ_{it}^{lsdv} + (v_{it} - \bar{v}_{i.}) \qquad RSS = 148,79$

Aus der obigen Schätzgleichung kann das für alle Wahlkreise identische Regressionsgewicht β direkt mit einem Wert von 3,14 abgelesen werden. Die individuell verschiedenen Regressionskonstanten der 4 Wahlkreise ergeben sich wie folgt:

$\alpha_1 = 34,15 - 3,14 \cdot 4,42 = 20,27$
$\alpha_2 = 46,32 - 3,14 \cdot 5,24 = 29,87$
$\alpha_3 = 24,89 - 3,14 \cdot 2,84 = 15,97$
$\alpha_4 = 45,16 - 3,14 \cdot 6,92 = 23,43$

Wie erwähnt, kann das "LSDV"-Modell auch durch die Einführung von Dummy-Variablen geschätzt werden, so daß die obige Mittelwertsbereinigung "automatisch" erfolgt. In unserem Fall gehen neben den untransformierten Beobachtungswerten, die aus Gründen der Bequemlichkeit noch einmal in den Spalten (3) und (8) aufgelistet sind, 4 entsprechend kodierte 0-1 Variablen in die Regressionsfunktion ein, die sich in den Spalten (4)-(7) befinden. Der Vorteil dieser Vorgehensweise ist, daß die individuellen Regressionskonstanten nun (bis auf vernachlässigbare Rundungsfehler) direkt aus der Regressionsfunktion ersichtlich sind

$$SPD_{it} = 20,29 \cdot D_{1t} + 29,88 \cdot D_{2t} + 16 \cdot D_{3t} + 23,45 \cdot D_{4t} + 3,14 \cdot ALQ_{it} + v_{it} \qquad RSS = 148,79$$

Da aber bei einer großen Anzahl von Untersuchungseinheiten die Einführung von gleich vielen Dummy-Variablen erforderlich ist, kann dies zu numerischen Problemen bei den Statistik-Programmen führen. Des weiteren, wie weiter unten gezeigt wird, bildet die Mittelwertberechnung den Ausgangspunkt der Schätzung des "EC"-Modells, so daß die Schätzung der Dummy-Variante nur bei einer geringen Anzahl von Individuen sinnvoll ist.

Um zu überprüfen, ob die Einführung von unterschiedlichen Regressionskonstanten für die 4 Wahlkreise sinnvoll war, berechnen wir die F-Statistik, in dem wir einen Vergleich der Abweichungsquadratsumme von "Classical Pooling"- und "LSDV"-Modell, korrigiert durch die Freiheitsgrade, vornehmen. Es ergibt sich eine empirische F-Statistik von

$$F_{emp.} = \frac{(508,26 - 148,79)/3}{148,79/15} = 12,08$$

Demnach, da der Wert von 12,08 deutlich größer ist als die kritische F-Statistik von 5,42 auf einem 1% Signifikanzniveau, können wir die Nullhypothese einer gleichen Regressionskonstanten nicht annehmen, und ziehen das "LSDV"- dem "Classical Pooling"-Modell gegenüber vor.

Als nächster Schritt wäre noch die Möglichkeit in Betracht zu ziehen, das "Error Components"-Modell zu schätzen. Wiederum ist hier eine Transformation der Beobachtungswerte vorzunehmen, für die wir aber schon einige Vorarbeiten geleistet haben. Wie in Abschnitt III.2b ausgeführt, ist für die Schätzung des "EC"-Modells der Faktor θ zu bestimmen, der sich als das Verhältnis zweier Standardabweichungen ergibt. Die "within"-Schätzung des "LSDV"-Modells hat uns schon die erste Varianz geliefert, indem wir die angegebene Abweichungsquadratsumme der Residuen durch die Zahl der Freiheitsgrade dividieren, so daß wir $\hat{\sigma}_v^2 = 148,79/15 = 9,92$ erhalten.

Um zur Varianz $\hat{\sigma}_1^2$ zu gelangen,ist die Durchführung der "between"-Schätzung notwendig. Diese entspricht einer Regression über die schon bestimmten 4 zeitlichen Durchschnitts-

werte der Wahlkreise von SPD-Anteil und Arbeitslosigkeit. Es ergibt sich folgende Regressionsfunktion

$$\overline{SPD}_{i.} = 11,83 + 5,31 \cdot \overline{ALQ}_{i.} + \overline{w}_{i.} \qquad RSS = 61,24$$

wobei die Varianz dieser Schätzung sich wiederum aus dem Verhältnis von Abweichungsquadratsumme und Freiheitsgraden ($N - K - 1 = 4 - 1 - 1 = 2$) ergibt, in diesem Fall $\hat{\sigma}^2_{\overline{w}} = 61,24/2 = 30,62 = \hat{\sigma}^2_1 / T$. Da die Varianz dieser Schätzung mit *T* multipliziert die gesuchte Varianz $\hat{\sigma}^2_1$ ergibt, erhalten wir den Wert 153,1 und für das Verhältnis θ errechnet man $\hat{\theta} = 1 - \sqrt{9,92/153,1} = 0,75$. Im letzten Schritt müssen die Beobachtungswerte der Wahlkreise wie im "LSDV"-Modell von ihren Mittelwerten bereinigt werden, allerdings zieht man im "EC"-Modell nur einen Teil, gegeben durch den Faktor $\hat{\theta} = 0,75$, von den ursprünglichen Werten ab. Für den Wahlkreis 1 ist diese Berechnung exemplarisch wiedergegeben:

$$y^{ec}_{11} = 38,46 - 0,75 \cdot 34,15 = 12,85 \qquad x^{ec}_{2,11} = 4,32 - 0,75 \cdot 4,42 = 1,01$$
$$y^{ec}_{12} = 35,32 - 0,75 \cdot 34,15 = 9,71 \qquad x^{ec}_{2,12} = 5,86 - 0,75 \cdot 4,42 = 2,55$$
$$y^{ec}_{13} = 30,78 - 0,75 \cdot 34,15 = 5,17 \qquad x^{ec}_{2,13} = 3,85 - 0,75 \cdot 4,42 = 0,54$$
$$y^{ec}_{14} = 35,34 - 0,75 \cdot 34,15 = 9,73 \qquad x^{ec}_{2,14} = 4,08 - 0,75 \cdot 4,42 = 0,77$$
$$y^{ec}_{15} = 30,83 - 0,75 \cdot 34,15 = 5,22 \qquad x^{ec}_{2,15} = 3,99 - 0,75 \cdot 4,42 = 0,68$$

Die transformierten Werte auch für die 3 anderen Wahlkreise finden sich in Tabelle 4. Mit Hilfe dieser transformierten Daten kann die Schätzung des "EC"-Modells erfolgen, es ergibt sich die untenstehende Schätzgleichung:

$$SPD^{ec}_{it} = 20,77 + 3,47 \cdot ALQ^{ec}_{it} + w_{it} \qquad RSS = 178,63$$

Das hier präsentierte Rechenbeispiel sollte nur der Veranschaulichung der allgemeinen Vorgehensweise dienen, sowohl *N* als auch *T* sind hier zu klein gewählt, um eine sinnvolle Inferenz oder Entscheidung zwischen den einzelnen Modellen zu ermöglichen. Trotzdem soll exemplarisch die Berechnung der ***Hausman***-Statistik zur Unterscheidung von "LSDV"- und "EC"-Modell vorgestellt werden, da sie nichts weiter als die Schätzung der obigen "EC"-Regressionsfunktion erfordert, die um die transformierte Reihe der Arbeitslosigkeit des "LSDV"-Modells erweitert wird. Es läßt sich die folgende Schätzfunktion ermitteln:

Tabelle 4: Transformierte Beobachtungswerte von "LSDV"- und "EC"-Modell

(1) y_{it}^{lsdv}	(2) $x_{2,it}^{lsdv}$	(3) y_{it}	(4) D_{1t}	(5) D_{2t}	(6) D_{3t}	(7) D_{4t}	(8) $x_{2,it}$	(9) y_{it}^{ec}	(10) $x_{1,it}^{ec}$	(11) $x_{2,it}^{ec}$
4,31	-0,1	38,46	1	0	0	0	4,32	12,85	0,25	1,01
1,17	1,44	35,32	1	0	0	0	5,86	9,71	0,25	2,55
-3,37	-0,57	30,78	1	0	0	0	3,85	5,17	0,25	0,54
1,19	-0,34	35,34	1	0	0	0	4,08	9,73	0,25	0,77
-3,32	-0,43	30,83	1	0	0	0	3,99	5,22	0,25	0,68
-0,8	-0,11	45,52	0	1	0	0	5,13	10,78	0,25	1,2
2,39	-0,47	48,71	0	1	0	0	4,77	13,97	0,25	0,84
0,69	-0,56	47,01	0	1	0	0	4,68	12,27	0,25	0,75
4	2,12	50,32	0	1	0	0	7,36	15,58	0,25	3,43
-6,27	-0,96	40,05	0	1	0	0	4,28	5,31	0,25	0,35
-2,03	-0,45	22,86	0	0	1	0	2,39	4,19	0,25	0,26
-6,37	-1,07	18,52	0	0	1	0	1,77	-0,15	0,25	-0,36
-1,96	0,02	22,93	0	0	1	0	2,86	4,26	0,25	0,73
0,13	0,31	25,02	0	0	1	0	3,15	6,35	0,25	1,02
10,24	1,17	35,13	0	0	1	0	4,01	16,46	0,25	1,88
-3,3	-0,43	41,86	0	0	0	1	6,49	7,99	0,25	1,3
-0,83	-0,74	44,33	0	0	0	1	6,18	10,46	0,25	0,99
-1,95	-0,87	43,21	0	0	0	1	6,05	9,34	0,25	0,86
1,53	0,85	46,69	0	0	0	1	7,77	12,82	0,25	2,58
4,54	1,2	49,7	0	0	0	1	8,12	15,83	0,25	2,93

$$SPD_{it}^{ec} = 11,86 + 5,3 \cdot ALQ_{it}^{ec} - 2,16 \cdot ALQ_{it}^{lsdv} + w_{it}^{*} \qquad RSS = 167,92$$

Dementsprechend ergibt sich eine empirische F-Statistk von

$$F_{emp.} = \frac{178,63 - 167,92 / 1}{167,92 / 17} = 1,08.$$

Da dieser Wert den kritischen F-Wert von 8,40 auf einem 1% Signifikanzniveau nicht übersteigt, kann die Nullhypothese nicht abgelehnt werden. Dies bedeutet, daß man in unserem Beispiel das "EC"- dem "LSDV"-Modell gegenüber vorziehen sollte.

Zusammenfassend läßt sich also festhalten, daß die hier vorgestellten Modelle zur Regressionsanalyse mit Paneldaten im wesentlichen nur die Verwendung von transformierten Beobachtungswerten erfordern. Die Transformationen selbst können recht bequem in Excel durchgeführt werden, da dieses Programm z.B. unter dem Menü-Punkt "Formel" und dem Unterpunkt "Funktion einfügen..." die Möglichkeit bietet, die Berechnung von Mittelwerten für alle Variablen en bloc vorzunehmen. Nachdem die Daten entsprechend umgerechnet und in SPSS eingelesen worden sind, kann anhand einer gewöhnlichen Regressionsanalyse die Schätzung der Parameter erfolgen. Einschränkend soll jedoch angemerkt werden, daß in diesem Beitrag auf wichtige Abweichungen von den Annahmen des klassischen linearen Regressionsmodells nicht eingegangen wurde, so wurde beispielsweise durchgängig von meßfehlerfreien Beobachtungswerten ausgegangen. An dieser Stelle sei allerdings auf die unten angeführte Literatur verwiesen.

Literatur:

Arminger, G. und ***F. Müller*** 1989:
Lineare Modelle zur Analyse von Paneldaten, Opladen: Westdeutscher Verlag.

Baltagi, B.H. 1995:
Econometric Analysis of Panel Data, New York: Wiley.

Dielman, T.E. 1989:
Pooled Cross-Sectional and Time Series Data Analysis, New York/Basel: Marcel Dekker.

Engel, U. und ***J. Reinecke*** 1994:
Panelanalyse: Grundlagen, Techniken, Beispiele, Berlin: de Gruyter

Hsiao, C. 1986:
Analysis of Panel Data, Cambridge: Cambridge University Press.

Johnston, J. 1986:
Econometrics Methods, 3rd ed., New York: McGraw-Hill.

Judge, G. et al. 1985:
The Theory and Practice of Econometrics, 2nd ed., New York: Wiley.

Judge, G. et al. 1989:
Introduction to the Theory and Practice of Econometrics, 2nd ed., New York: Wiley.

Kmenta, J. 1986:
Elements of Econometrics, 2nd ed., New York: Macmillan.